学校编码：10384　　分类号_____密级 _____
学　　号：200431047　　UDC_____

厦门大学

硕 士 学 位 论 文

基于终端的 EST 分析方法及平台构建

A New Data Processing Method & System Construction for Expressed Sequence Tags Based on cDNA Termini

刘 元 生

指导教师姓名：吉国力 教授
梁 春 博士

专 业 名 称：系 统 工 程
论文提交日期：2008 年 6 月
论文答辩日期：2008 年 月
学位授予日期：2008 年 月

答辩委员会主席：_____

评　　阅　　人：_____

2008 年 6 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　月　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1.保密（），在年解密后适用本授权书。

2.不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

摘　要

本文是和美国迈阿密大学植物系梁春博士（Dr. Chun Liang）合作，对 EST 序列的处理方法及平台构建进行深入研究和探索后得到的成果。目前，已经有许多其它 EST 序列预处理软件广泛的应用于 EST 序列的处理之中，但是目前所有这些 EST 序列预处理软件都有一个共同的缺点，就是忽略了 cDNA 的文库构造过程中所形成的终端及其结构信息，导致目前经过这些 EST 序列预处理软件得到的相当一部分 EST 序列存在信息污染、信息缺失、数据不完整及数据不一致等一系列问题，这些存在以上问题的 EST 数据源将影响到 EST 序列的后续分析和处理，甚至得出错误的结论。EST 序列中的终端特征是标识 cDNA 片段在 EST 序列中的位置的重要信息，且终端本身包含某些生物学现象，因此识别出 EST 序列中的终端特征有助于生物学家发现其中的变化形式及其生物学意义。

本论文提出了一种基于终端的 EST 序列处理方法，根据 cDNA 文库构造过程中所形成的终端及其结构信息，利用序列比对算法识别出 EST 序列中的终端特征，并根据 EST 序列中的终端、终端结构及其它信息提取干净序列，这种处理方法能给生物学家提供一个更干净、更可靠和更完整的 EST 数据源。论文在基于终端的基础上，分析和介绍了一套完整的 EST 序列处理流程及其后续分析，并对如何构建 EST 序列分析平台及其相关技术进行了深入分析和探讨。论文最后，将基于终端的 EST 序列处理方法，应用到莱茵衣藻 EST 序列处理中，在该实例中，首先对 EST 序列的终端特征进行统计，引出终端及干净序列提取的复杂性，论文接着通过各种统计数据，说明和验证了基于终端的 EST 序列处理方法的正确性和可靠性，且这种分析方法确实能给生物学家提供更干净、更可靠和更完整的 EST 数据源，为生物学家今后处理 EST 序列提供一种全新的处理方法。

关键词：生物信息学；EST 预处理；序列比对

ABSTRACT

Based on the cooperation with Dr. Chun Liang at the Department of Botany, Miami University, USA, this thesis is focus on EST data preprocessing and system construction. Currently, there are a lot of pipelines largely used in EST data preprocessing, but all of these pipelines doesn't handle the cDNA termini and its structures which are formed in the cDNA library construction, because of this, many sequences, which are processed by these pipelines, have the problem of contamination, these EST sequences contained errors exert deleterious impacts on downstream EST-based applications. Moreover, cDNA termini are important information used to annotate the boundary of cDNA fragment in the EST sequence, and cDNA termini itself represent some important biological phenomenon, so it is important to identify the cDNA termini in EST sequences for further research conducted by biologists.

In this thesis, we imported a unique method to process the EST sequences, according to the cDNA termini and its structures formed in cDNA library construction, we identify cDNA termini in EST sequences using sequence alignment algorithm, and then we fetch the clean cDNA fragments from EST sequences according to the identified termini, termini-structures and other relevant features. Using this unique method, we provide the biologists a new EST data stream with higher quality. Based on cDNA termini, we introduced and analyzed an integral data processing pipelines for EST sequences in this thesis, and we also presented some detail information about how to construct the system for EST data processing and visualization. In the last of this thesis, we used this unique EST data processing method to process the EST sequences of Chlamydomonas reinhardtii organism, in this application, we analyzed the complexity of cDNA termini and its structures, and then other statistic results from different views are used to verify the correctness of EST data processing method based on cDNA termini. From the analysis results, we are sure that this new EST data processing method can provide the biologists a more accurate and stable data stream, and provide the biologists another choice in EST data processing.

Key words: Bioinformatics; EST Preprocess; Sequence Alignment.

目　　录

Contents

第一章 概述

1.1 背景

EST 即“Expressed Sequence Tag"，中文称作“表达序列标签”，具体来讲，EST 来源于特定条件下某种组织在某一时期的 cDNA 文库，然后从其中随机挑选一个 cDNA 克隆体进行 5'端和 3'端一次性测序获得的短的 cDNA 序列片段，一般长度为 400-600bp，代表了一个完整基因的一部分。从 1991 年 EST 这个概念被 Adams 及其同事提出以来[1]，EST 序列数据的积累异常迅速，现已成为 NCBI 收录的序列信息中最多的一种，截至 2008 年 2 月 1 日，NCBI 的 dbEST 数据库中已收录 EST 序列 49,693,316 条。

巨量积累的 EST 序列是一种宝贵的生物信息资源，在标注基因结构、基因功能、基因表达的变化形式、SNP 位点及预测和标注 polyA 剪切位点[2][3][4][5]等方面均具有重要应用价值，而干净的 EST 数据源是这些应用研究的基础。但是，目前所有的 EST 序列预处理软件，都忽略了 cDNA 文库构造过程中所形成的终端及其结构信息，导致经过这些 EST 序列预处理软件得到的 EST 数据源中存在信息污染和信息缺失等现象[6][7][8]，这些存在污染和信息缺失的 EST 数据源会影响到 EST 序列的后续分析，甚至推导出错误的结论；终端是标识 cDNA 片段在 EST 序列中的位置的重要信息，且终端本身包含某些生物学现象，标注 EST 序列中的终端信息，能帮助生物学家研究和分析基因的转录过程及其变化形式。

基于此，本文提取了一种新的 EST 序列分析方法，既基于终端的 EST 序列分析方法。

1.2 与论文相关的一些定义

脱氧核糖核酸[9]（DNA）

脱氧核糖核酸（deoxyribonucleic acid， DNA）分子是除一些病毒外所有生物体的遗传信息载体。DNA 是一种包含有核苷酸单体链的聚合物。每个核苷酸包括一个五碳糖，一个碱基和一个磷酸基。有四种类型的碱基，腺嘌呤、胞嘧啶、鸟嘌呤、胸腺嘧啶（分别用符号 A、C、G、T 表示），划分为嘌呤（purine）和

嘧啶（pyrimidine）两组。

DNA 具有双螺旋结构。它是由互补的两条链组成。两条链的结合遵循互补碱基配对法则，即嘌呤只能与嘧啶配对，反过来也一样，嘧啶只能与嘌呤配对，既碱基 A 与碱基 T 配对，碱基 C 与碱基 G 配对，这样产生了一个对称的双螺旋结构。

核糖核酸[9]（RNA）

另一类核酸称为核糖核酸（ribonucleic acid，RNA），它与 DNA 在以下几方面略有不同：

(1) 细胞的 RNA 是单链的；DNA 是双链的。

(2) RNA 含有核糖，而不是在 DNA 中发现的脱氧核糖。

(3) RNA 含有尿嘧啶（U）而不是胸腺嘧啶（T），且 U 与 A 配对。

(4) RNA 分子比 DNA 分子短得多。

RNA 主要在蛋白质合成中起作用，大量的科学实验表明，信息的传递不是由 DNA 直接传递给蛋白质的，而是在细胞核中先把 DNA 的遗传信息传递给 RNA，然后 RNA 进入细胞质中，在蛋白质合成中起模板作用。我们把这种 RNA 形象地叫做**信使 RNA（messenger RNA， mRNA）**。

染色体[9]（chromosome）：遗传物质——DNA 的携带者，不同的生物体具有不同的结构和数目。每条染色体又含有若干条基因，基因经过转录形成 mRNA，部分 mRNA 经过翻译形成蛋白质，也有一部分 mRNA 对基因的转录和 mRNA 的翻译起着调控作用。

基因：是指携带有遗传信息的 DNA 或 RNA 序列，也称为遗传因子，是控制性状的基本遗传单位。基因通过指导蛋白质的合成来表达自己所携带的遗传信息，从而控制生物个体的性状表现。基因是携带遗传信息的 DNA 片断，它携带编码蛋白质的信息。基因分为非编码区和编码区，编码区域称之为**外显子**，而间隔外显子之间的非编码区域称之为**内含子**，当基因转录成 mRNA 的时候，基因中的内含子被切除，而外显子得到保留，实际上真正编码蛋白质的是外显子，而内含子则无编码功能，内含子的具体功能还不清楚，可能在剪接过程中起作用，也可能是一些曾经有用但现在被弃用的序列。

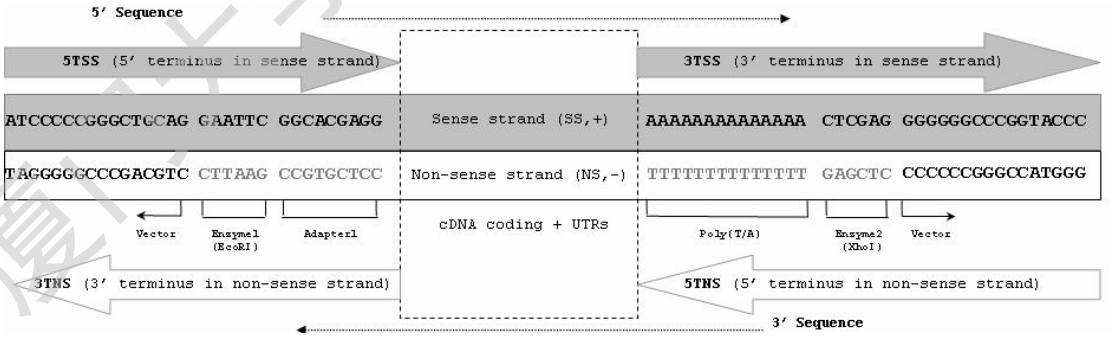
cDNA[9]（complementary DNA）：互补 DNA，以 mRNA 为模板在反转录酶作用下产生的产物名称。与基因组库一样，cDNA 库也是指一群含重组 DNA 的

细菌或噬菌体克隆。每个克隆只含一种 mRNA 的信息，足够数目克隆的总和包含细胞的全部 mRNA 信息，这样的克隆群体就叫做 cDNA 文库。与基因组文库相比，cDNA 文库便于克隆和大量扩增，可以从 cDNA 文库中筛选到所需目的基因，并用于该目的基因的表达。cDNA 文库是发现新基因和研究基因功能的基础工具。

限制性内切酶 [10]（Restriction Enzyme）： 20 世纪 60 年代末，在大肠杆菌中首先发现了一种酶，它会准确识别外来的 DNA，并且在特定的位点把后者切断，这就是限制性内切酶，在本论文中简称为限制酶。许多限制性内切酶具有回文特性，既它们在 DNA 双螺旋的两条链上等同。当剪切点在两条链上错开，形成“粘端”，就更有利于加工后重新联接，因而在基因工程中有广泛的应用。

EST（expressed sequence tag）：表达序列标签，是从已建好的 cDNA 文库中随机取出一个克隆，对插入的 cDNA 片段进行 5’或 3’端一轮单向自动测序，所获得的一段序列。

EST 终端（terminus）:是用于标识 cDNA 片段在 EST 序列中的位置的一段字符序列，EST 终端的形成与 cDNA 文库的构造过程及克隆载体有密切联系，由于 DNA 是互补的双螺旋结构，因此在 EST 的两个测序方向上各存在一个开始终端和结束终端，图 1 为莱茵衣藻（Chlamydomonas reinhardtii）EST 序列终端结构。



图片来源：http://www.conifergdb.org/chlamyest/terminiStructureInfo.php

图 1 莱茵衣藻 EST 序列终端结构

在上图中 5TSS 表示 5’方向 EST 序列的开始终端，3TSS 表示 5’方向 EST 序列的结束终端，5TNS 和 3TNS 分别是 3’方向 EST 序列的开始和结束终端，并不是每条 EST 序列都包含开始终端和结束终端，由于目前测序技术水平对 EST 序列的长度有一定限制，因此大部分 EST 序列并不包含结束终端。终

端及其结构的具体定义和 cDNA 文库的构造过程及克隆载体的选择有密切的关系，在莱茵衣藻 EST 序列中，本文给出的终端的具体定义如下：

(a) **5TSS**：是 5’测序方向的开始终端，依次由载体片段、限制酶(ECORI：GAATTC)和 adaptor(GGCACGAGG)组成。

（b）**3TSS**：是 5’测序方向的结束终端，依次由 polyA、限制酶（XHOI：CTCGAG）和载体片段组成，其中 polyA 长度至少为 8。

(c) **5TNS**：是 3’测序方向的开始终端，依次由载体片段、限制酶（XHOI：CTCGAG）和 polyT 组成，其中 polyT 长度至少为 8。

（d）**3TNS**：是 3’测序方向的结束终端，依次有 adaptor(CCTCGTGCC)、限制酶(ECORI：GAATTC)和载体片段组成。

1.3 EST 技术概要

1.3.1 EST 序列的产生过程

（1）构建 cDNA 文库[11]

EST 是从 cDNA 文库中随机挑选的一条 cDNA 进行克隆得到的序列片段，cDNA（complementary DNA）既互补 DNA，是以 mRNA 为模板在反转录酶作用下产生的产物名称。cDNA 的产生包括以下几个关键过程（如图 2 所示）：

(a) 转录（transcription）

基因 DNA 含有基因序列以及调控序列，在一定机制和条件下，基因序列被转录出来成为 pre-mRNA。

(b) 剪切（splicing）和多聚腺苷化(polyadenylation)

通过上一步形成的 pre-mRNA 含有内含子(intron)和外显子(exon)，经过剪切这个步骤，可以将 pre-mRNA 的内含子序列移去。同时 mRNA 的 5'端加上甲基化鸟嘌呤帽子。在 pre-mRNA 3'末端由于 RNA 内切酶的作用，在特定的位点上切割掉一部分，形成新末端。在这新末端上，一串腺嘌呤核苷酸，称为多聚 A 尾巴(poly(A) tail)需加到 3'末端，最后形成成熟的 RNA(mRNA)。

(c) 反转录，互补合成 cDNA

在生物体内，mRNA 中启始密码子(start codon)和终止密码子(stop codon)之间的编码区域(coding segment)经过翻译表达就合成了蛋白质。而如果 mRNA 在体

外用逆转录酶、DNA 合成酶等处理就可以合成双链 DNA，并在两端形成与载体酶切点可以相连的接头，这就是双链 cDNA。

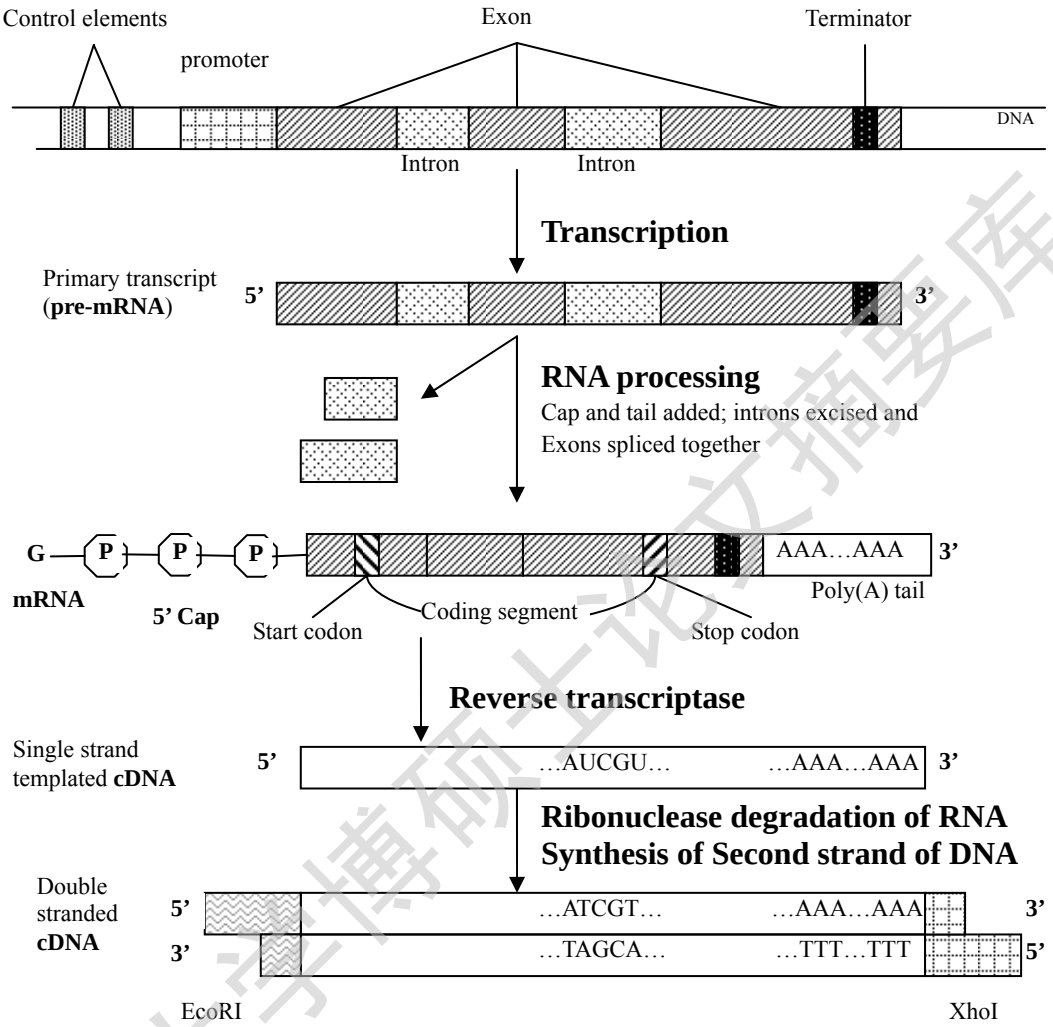


图 2 cDNA 产生过程

（2）克隆 cDNA

通过上一步，从 mRNA 转化而来的双链 cDNA 形成了可以与载体酶切点相接的接头，然后再选择合适的载体，一般是改造过的质粒，如加载多克隆位点接头片段、并把接头片段两侧区域改为特定的测序引物序列，将双链 cDNA 接入载体中进行大量的克隆。

（3）EST 测序

现在的大规模自动测序基本上都是基于 Sanger 的“DNA 双脱氧链末端终止测序法”进行的。具体过程是：先从文库中随机挑取大量克隆，在体外变性为单链后，利用多克隆位点接头两侧序列设计载体通用引物(T3 对 5'端，T7 对 3'端)

进行两端一次性自动化测序，可以测出 400-600bp 的序列。

1.3.2 传统的 EST 序列处理流程

将 EST 序列用于基因标注及后续分析之前，必须对 EST 序列进行处理，包括 base calling、去除载体和重复序列、过滤序列的低质量区、聚类拼接，图 3 是 EST 序列的处理流程。

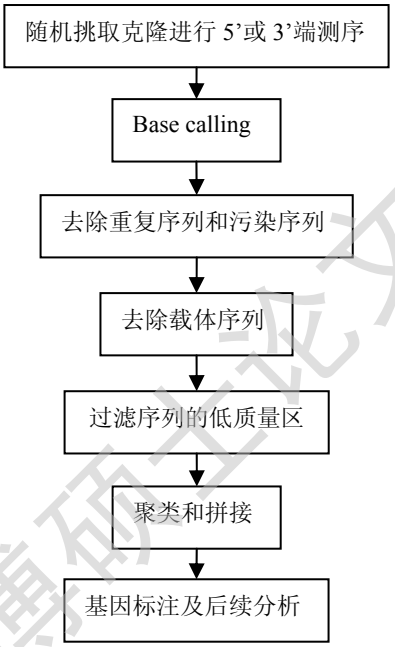


图 3 EST 序列处理流程

（1）Base Calling

通过机器测序得到的是序列的色谱文件（trace file），不适合计算机进行大规模的计算和处理，因此需要通过 Base calling 软件将色谱文件转换成适合计算机处理的碱基序列，当前 dbEST 数据库中的大部分 EST 数据都是用 phred 软件来做 Base calling，序列的色谱文件经过 phred 软件处理过后，会生成碱基字符序列，并且对每个碱基生成一个质量值，用来评估该碱基的可靠性[12][13]。

（2）去除重复序列和污染序列

EST 序列中可能包含不属于表达基因的假序列，如重复序列、污染序列（核糖体 RNA、细菌或其它物种的基因组 DNA 等），这些重复序列和污染序列会影响 EST 序列的后续分析。可以使用 RepeatMasker、MaskerAid、Dust 等软件去除包含重复序列和污染序列的 EST。